

AudioPaLM: A Large Language Model That Can Speak and Listen

Year: 2023 | Citations: 378 | Authors: Paul K. Rubenstein, Chulayuth Asawaroengchai, D. Nguyen

Abstract

We introduce AudioPaLM, a large language model for speech understanding and generation. AudioPaLM fuses text-based and speech-based language models, PaLM-2 [Anil et al., 2023] and AudioLM [Borsos et al., 2022], into a unified multimodal architecture that can process and generate text and speech with applications including speech recognition and speech-to-speech translation. AudioPaLM inherits the capability to preserve paralinguistic information such as speaker identity and intonation from AudioLM and the linguistic knowledge present only in text large language models such as PaLM-2. We demonstrate that initializing AudioPaLM with the weights of a text-only large language model improves speech processing, successfully leveraging the larger quantity of text training data used in pretraining to assist with the speech tasks. The resulting model significantly outperforms existing systems for speech translation tasks and has the ability to perform zero-shot speech-to-text translation for many languages for which input/target language combinations were not seen in training. AudioPaLM also demonstrates features of audio language models, such as transferring a voice across languages based on a short spoken prompt. We release examples of our method at <https://google-research.github.io/seanet/audiopalml/examples>